Proceedings Paper:
Alrashdi, Reem and O'Keefe, Simon orcid.org/0000-0001-5957-2474 (2020) Automatic Labeling of Tweets for Crisis Response Using Distant Supervision. In: WWW '20: Companion Proceedings of the Web Conference 2020. ACM , pp. 418-425.

https://doi.org/10.1145/3366424.3383757

Automatic Labeling of Tweets for Crisis Response Using Distant Supervision

Reem ALRashdi
Computer Science Department, University of York, York, UK, CSSE Department, University of H'ial, H'ail, SA
rmma502@york.ac.uk

Simon O'Keefe
Computer Science Department, University of York, York, UK
simon.okeefe@york.ac.uk

ABSTRACT

Current tweet classification models aimed at enhancing crisis response are based on supervised deep learning. They rely on the quality and quantity of human-labeled training data. Still, the available training data is small in size and imbalanced in coverage of crisis types, which prevents the models from generalization, and as it is manually labeled, it is also expensive to produce. To overcome these problems, distant supervision can be applied to automatically generate large-scale labeled data for tweet classification for crisis response. Experimental results on different crisis events show that our work can produce good quality labeled data from past and recent events. Substituting automatically labeled training data for part of the manually labeled training data has a minimal impact on the model performance, indicating that automatically labeled data can be used when no hand-labeled data is available.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Natural language processing**; • **Information extraction**;

KEYWORDS

Tweet classification, crisis response, distant supervision, large- scale data, FrameNet

ACM Reference Format:
Reem ALRashdi and Simon O'Keefe. 2019. Automatic Labeling of Tweets for Crisis Response Using Distant Supervision. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3366424.3383757

1 INTRODUCTION

During crises, people spread the news on Twitter and share valuable, real-time and on-topic information like their statuses, injured or dead people and the damage caused by the crises [1]. They also tweet to ask for help or offer help to others. Because of that, Twitter has become a dominant platform for organizations and people to post or gather information in many natural or human-made crises during recent years [2], such as earthquakes [3], floods [4], wildfires [5] and nuclear disasters [6]. For example, in 2011, 177 million crisis-related tweets were published in only one day during an earthquake in Japan [7].

Situational awareness can be significantly enhanced by people-generated tweets [1]. These tweets can be used by large-scale disaster response organizations to make better decisions and quick responses. However, humanitarian organizations cannot manually observe, process and convert the enormous volume of data into actionable information [8]. Thus, they do not widely use social media data such as Twitter in their disaster response operations [9].

Tweet classification for crisis response is a text classification task that aimed at identifying whether a tweet is related to a specific crisis event or not. For example, " BREAKING: Nepal police official says at least 1,910 have died, including 721 in Kathmandu, in the quake " is a tweet related to a Nepal Earthquake event while "So important! Hindu, Buddhist, Christian and Muslim leaders denounce #childmarriage in joint broadcast in Nepal" is irrelevant. The main purpose of binary tweet classification models is to reduce the volume of tweets in real- time to simplify the work for humanitarian organizations to respond to people in need in the early hours of a crisis. However, current tweet classification models suffer from the lack of labeled data [10], which prevents them from reaching a generalized model [11] as tweets related to various crisis types have different features and social media response [12]. Besides, it is infeasible to manually annotate tweets for every crisis event, especially in real- time [13]. Because of that, semi-supervised approaches that automatically generate labeled training data from an unlabeled corpus are desirable.

The authors in [14] applied semi-supervised learning to disaster response data. They employed self-training learned on the small available data to label a new collection, Mayanmar _Earthquake_2016, derived from Twitter with annotation accuracy of 80%. Similar to [14], we build a semi-supervised method, but we use distant supervision [15] via an external linguistic knowledge base, FrameNet [16], instead of self-labeling. Unlike [14], we propose a novel framework that does not duplicate the label noise exists in the current dataset and explores different ranges of unseen features by expanding the original keyword list to include new linguistic units (new keywords with similar meaning) derived from FrameNet which provides the chance to improve the generalization level of the classification model.

1.1 Contribution

Our work addresses the problem of the low generalization level of the crisis-related classifier caused by the lack of annotated tweets. To reduce the generalization error, we present a new framework to

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383757

label new crisis event data. This data will be added to the available ones to train the classifier to filter the massive volume of tweets posted by people during crises. Our main goal is to investigate, for the first time, the application of distant supervision in producing good-quality labeled-data for our task. Specifically, our research questions are:

- Can we automatically generate labeled training data for tweet classification for crisis response that has a competitive quality level compared to manually labeled training data.
- When added to the available labeled data, does our automatically generated labeled data improve the performance of the crisis-related tweet classifier.

Also, and to evaluate our work, we create a new collection of crisis-related labeled examples from Twitter data from new disaster events: 2018 Texas Floods, 2018 Indonesia Earthquake and 2018 Sunda Strait tsunami.

The rest of the paper is organized as follows: The following section covers the related works in tweet classification for crisis response and distant supervision fields. Next, our proposed labeling framework is described in detail. After that, our experimental setup and results are discussed. In the last section, we provide a conclusion of our investigation and suggestions for future work.

2 RELATED WORK

2.1 Tweet Classification for Crisis Response

In the literature, many tweet classification methods have been introduced to reduce the enormous volume of tweets posted by people during crises to simplify humanitarian organizations' work. These methods rely on two main approaches: matching-based and learning bases approaches [17].

2.1.1 Matching-based approach. The purpose of this approach is to identify the related tweets based on predefined keywords and hashtags [18]. The authors in [19] built CrisisLex based on keywords and hashtags related to the crisis events. This method, however, unable to retrieve related tweets do not contain these keywords or hashtags even if the tweets contain words with similar meanings. Another issue is that they mislabel unrelated tweet mention one of the hashtags or keywords where no noise reduction technique is used. Geolocation has also been used as a feature to retrieve related tweets; however, this feature does not exist for most of the posted tweets [20]. To solve these issues, platforms such as CrisisTracker have been developed to enable humans to label disaster-related tweets [21]. Yet, this method is costly because it requires a lot of time, money and effort to manually label a large number of crisis events from different locations and circumstances.

2.1.2 Learning-based approach. Unlike the matching-based method, studies in the learning-based approach aim to build a model based on a set of labeled tweets from crisis events to identify crisis-related tweets from unseen examples. Recently, deep learning algorithms are proven to outperform tradition machine learning algorithms. In [2] and [10], the authors used Convolutional Neural Networks (CNN) to classify tweets for crisis response based on their relatedness to a given crisis event or their information type. Nevertheless, these models are known to suffer from low generalization levels when tested on the unseen (out-of-domain) disaster event. The reason behind that is the lack of manually labeled data available to train these deep learning models, especially that they heavily rely on the quality and the quantity of the training data [11, 14, 22]-[24]. Our work aimed at solving this unaddressed issue.

2.2 Distant Supervision

Since 2009, distant supervision has been successfully applied to label training data via an external knowledge base in many Natural Language Processing (NLP) tasks such as event extraction [25, 26], sentiment analysis [27], topic classification [28, 29] and relation extraction [30]. [25] employed distant supervision for event extraction using frames from FrameNet as event types and the linguistic units as triggers that evoke the event. [26] proposed a combination framework of a relational and a linguistic knowledge bases on Wikipedia data, Freebase and FrameNet, respectively.

In addition, distant supervision technique has been useful to label Twitter data with different approaches. In [27], the authors assumed that the emotions in the tweets express the feeling of the writers. This assumption has been used to label tweets for sentiment analysis tasks. For example, if the tweet contains happy face emotion, then the tweet is labeled as positive. [28] applied distant supervision to a topic classification task where they transfer labels from tweets of topically-focused Twitter accounts to tweets posted by general Twitter accounts. [29] used YouTube videos to assign labels to tweets containing links to these videos. However, people do not usually use emotions and YouTube videos when posting information during crises. Also, there is a lack of crisis-related Twitter accounts. On the other hand, keywords play a vital role in identifying disaster-related tweets in crisis situations. Thus, we use them in applying distant supervision to enhance the crisis response process. In our work, if one of the top crisis type keywords exists as a lexical unit of a frame in FrameNet, then distant supervision assumes that all the lexical units related to the given frame express the given crisis type.

Several selections, noise reduction and generating negative examples techniques have been introduced under the umbrella of distant supervision to achieve the best results.

2.2.1 Selection methods. Pointwise Mutual Information (PMI) is a well-known method to calculate the importance of a feature (keyword) in a given category (class) [31]. In the context of event detection, [14] used the mean PMI to select the most related features in the disaster lexicon. Moreover, KeyRate (KR) has been developed by [26] to select the most important triggers and arguments for a specific event type for event extraction tasks. In our work, we use a method inspired by [26]; however, we change some variables to suit the case of the binary classification task instead of the multiclassification task discussed in [26].

2.2.2 Noise Reduction. Noise is a recognized labeling problem in using distant supervision for labeling raw data. This problem can seriously affect the performance of deep learning models and hence has been well-addressed in the literature. For relation extraction task, a multi-instance single-label model has been introduced by [32]. They assumed that each entity pair holds at least one relation

expression. This work has been extended to a multi-instance multi-label model by [33], where more than one label is allowed for each entity pair. Besides, noise reduced in other works by other approaches. [34] filtered the noise in the positive examples by using a threshold for the frequency of the dependency paths among these examples. [35] and [36] applied three heuristic labeling methods that were initially proposed by [37]: top trigger words, closest pairs and high-confident patterns. In the event detection literature, [26] used two external knowledge bases instead of one to generate large scale distant supervision data. FrameNet has been used to eliminate the noisy trigger words and expand the trigger list to include new triggers. To filter the noise in our distant supervision data, we only take into account the tweets with two keywords from the final list instead of only one keyword. In addition, all the tweets contain only one keyword are ignored to reduce the noise reduction caused by using weak keywords from FrameNet.

2.2.3 Generating Negative Examples. The simplest way to generate negative examples is to apply the against assumption of distant supervision. For instance, if the distant supervision assumes that every sentence contains at least one existing pair in the external dataset, then this sentence expresses that relation and thus labeled as positive. In this case, negative examples can be generated directly when the sentence has no such pairs. However, applying this simple technique may cause a lot of noise in the labeled data. Several works have been introduced to avoid this problem. In our framework, we assume that the tweet with no keywords from the final list does not express the crisis type in any way, and thus we label them as negative tweets.

Our main goal in this research is to examine the application of using distant supervision in generating large scale labeled data with less amount of money, time and competitive quality compared to manually labeled data to fulfil the urgent need of more training data for crisis response. In this paper, we propose a framework to label unlabeled tweets from new crisis events retrieved using Twitter API.

3 METHOD

Our method is described by the following steps (as shown in Figure 1):

1. Creating the initial keyword list. The list is created based on the available annotated tweets from different collections related to the same crisis type. For example, all the available manually labeled data for all the Earthquake events are used to establish the initial keyword list for the crisis type Earthquake. The initial Earthquake keyword list in this step includes an unlimited number of words without any restrictions. To avoid word redundancy and reduce the amount of linguistically similar words, we use the Snowball Stemmer tool from NLTK 3.4 to stem each word to its root.

2. Selecting top K keywords. After extracting the initial list of crisis type keywords, the top K keywords are then chosen based on an intrinsic filtering method where a statistical measurement is used to pick the keywords with the highest scores. We calculate the Keyword (KW) value for each keyword in the initial keyword list. In a tweet, a word that describes a given crisis type can be a verb, a noun, or an adverb. For instance, "magnitude" (noun), "shake" (verb)

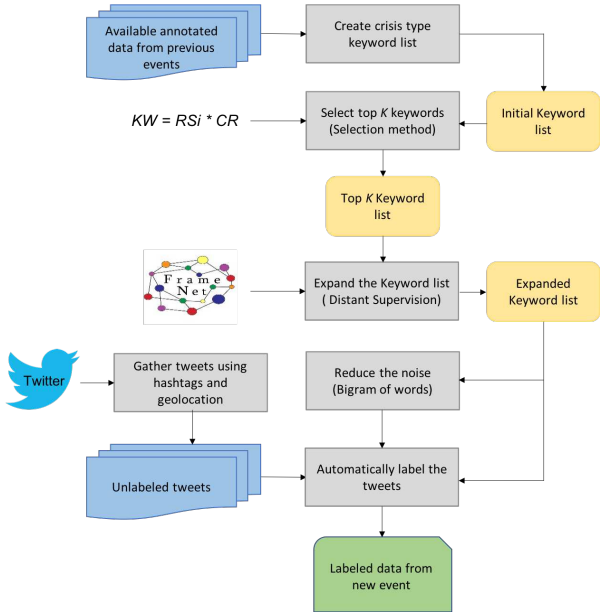

Figure 1: Distant supervision-based Framework.

and "deadly" (adv) can be considered to be keywords of the crisis type Earthquake. Intuitively, if the word appears more than other words in the labeled tweets of a given crisis type, then the word most likely describes this crisis type. If the same word appears in both positive and negative examples of the crisis type, then the word has a lower probability to describe the crisis type. Thus, KW is calculated as follows:

$$\mathrm{RS_i} = \frac{\mathrm{Count(W_i\ \ CT)}}{\mathrm{Count(CT)}} \tag{1}$$

$$\mathrm{CR_i} = \log \frac{3}{\mathrm{Count(\ CTC_i\ )}} \tag{2}$$

$$\mathrm{KW_i} = \mathrm{RS_i} * \mathrm{CR_i} \tag{3}$$

Where RS_i (Role Saliency) represents the saliency of i-th keyword to identify a specific word of a given crisis type, $Count(W_i, CT)$ is the number of a word W_i occurs in all the tweets related to the crisis type CT and $Count(CT)$ is the count of times all words occurring in all the tweets related to the crisis type. The KW equation is inspired by [26] where they used a similar Key Rate (KR) value to detect key arguments in event extraction tasks; however, unlike [26], CR_i (Crisis Relevance) in our work represents the ability of the i-th keyword to distinguish between the tweets related to the crisis type and irrelevant tweets, and $Count(CTC_i)$ equals 1 if the i-th keyword occurs only in the related tweets and 2 if the i-th keyword occurs in both related and irrelevant tweets. Finally, and after removing stop words such as "and," hashtags such as "#earthquake," places such as "Nepal" and useless twitter-specific words such as "RT" and "via," we compute KW_i for all the words in the initial keyword list from step one and sort them according to their KW values. At the end, we pick the top K keywords of a given crisis type. For example, for crisis type Earthquake, the top K words list contains "earthquake", "hit" and "magnitude," which have the highest KW values comparing to the other words in the initial Earthquake list

Table 1: KW values of words from Earthquake keywords list.

Keywords	KW values	Ranking
Help	0.00495	5
Quake	0.00702	3
Hit	0.00449	6
Kill	0.00216	25
Aftershock	0.00199	28
Give	0.00114	77
New	0.00129	62

from step one. The *KW* value for a given word increases when the *RS* or *CR* value of the same word increases. *RS* rises only if the frequency of the word in the related tweets rises. In contrast, *CR* increases in one case where the word occurs only in the related tweets. Table 1 shows how *KW* values play an important role in indicating the strongest keywords of Earthquake crisis type, where we can see that crisis-related and earthquake-related words have higher *KW* values than the unrelated ones.

Other methods, such as PMI or Term Frequency-Inverse Document Frequency (TF-IDF), have not been used here for solid reasons. PMI, where we calculate PMI for positive examples and PMI for negative examples to calculate the final PMI score, is not a fair metric in our case because of the imbalanced data problem given the limited available manually-labeled data where the number of positive examples is higher than the number of negative examples in all events as shown in Table 2. On the other hand, the imbalanced dataset problem does not affect our formula as $Count(CT)$ takes into account the total number of words in the related tweets only, while the total number of words in the unrelated tweets is ignored.

TF-IDF is another selection method that aims at measuring the importance of a feature (word) to a given document (event type) in a given corpus (collection of tweets) [38]. However, this selection method is not suitable for our case because IDF has more impact on the final result than TF, while, in our case, they should be equally important since tweets are short and full of noise. If we used TF-IDF on our data, rare words such as hashtags, mentions and misspelled words will have higher TF-IDF than essential keywords. Also, an important keyword may appear in both related and not related tweets. For instance, in Earthquake crisis type data, "earthquake" may appear very frequently in related Earthquake event tweets and once or twice in unrelated Earthquake event tweets. On the other hand, our method does not discard the impact of word frequency if the word appears in both related and unrelated tweets.

3. Applying distant supervision. The list contains *K* keywords is then expanded to include similar linguistic units from FrameNet. FrameNet is an external linguistic knowledge base for English that consists of more than 1000 semantic frames that have more than 100,000 Lexical Units (LU), lemmas and part of speech tags. Each frame in FrameNet is associated with a group of LUs that evoke that frame. In our work, these elements can be used as the keywords that evoke the crisis type. We retrieve all the LUs of a given frame if the crisis keyword is one of these LUs, and the frame is related to the crisis type. For example, "aid.v" is a linguistic unit related to the frame Assistance in FrameNet which is inherited

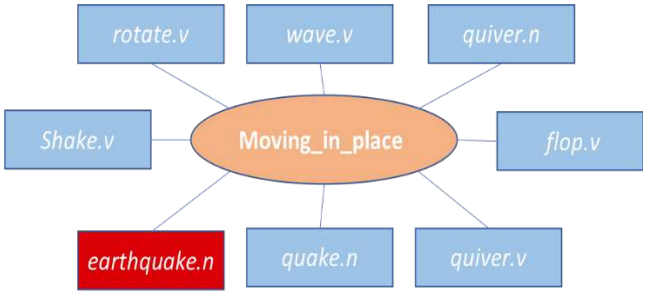

Figure 2: FrameNet example of Moving_in_place frame (in orange), its associated Lexical Units (in blue) and the keyword from our crisis type keyword list for Earthquake events where the LUs are mapped, earthquake.n (in red).

from Intentionally_act and can be mapped to "help" which is a crisis keyword gathered from the first step and has been picked in the second step as one of the top *K* keywords according to its high *KW* value. In other words, if one of the top crisis type keywords exists as a lexical unit of a frame in the database, then distant supervision assumes that all the lexical units related to the given frame express that crisis type. Figure 2 shows an example of a frame and its associated LUs and how we map them to keywords.

4. Retrieving unlabeled tweets from a new crisis event. Unlabeled tweets of a new crisis event are obtained from Twitter using the Twitter API. Hashtags are used as the initial indicator of the crisis-related tweets along with geolocation information of the crisis location. For example, we use the hashtags "#earthquake", "#prayforriyadh", "#riyadhearthquake", or any other widespread hashtags related to Riyadh Earthquake event and the geolocation of Riyadh city. Unlabeled tweets from multiple hashtags can also be merged. Although hashtags can be a beneficial method to classify related and unrelated tweets, there is still a considerable number of unrelated tweets where people use the same hashtag while tweeting about irrelevant subjects, such as advertising for a particular product or service. Moreover, this step can be seen as hashtag-based supervision where tweets may contain some of these topical hashtags. However, not all the hashtags are in the keyword list of the related crisis type. Also, in the previous steps, we replace all the words starting with the symbol "#" with the word hashtag, which eliminates any possible active role of the topical hashtags in the list.

5. Noise reduction. We filter the unlabeled tweet set gathered from step four by applying a specific lexical feature (bigram of keywords). After filtering the unlabeled tweets, only the examples with two keywords from the final keyword list remain. This step reduces the effect of a powerful hashtag where the hashtag without symbol # is one of the keywords. It also eliminates several tweets that contain only one weak keyword expanded from FrameNet, which decreases most of the noise caused by step three.

6. Labeling the corpus as related and not related examples. A collection of data from the given crisis event is automatically generated by labeling tweets from step five as relevant examples and tweets with no keywords from the expanded keyword list as not related (negative) examples.

Table 2: Summary of the collections used in our experiments from CrisisNLP, CrisisLex26 and CrisisLexT26. The abbreviations in the table represent the type of the data, the place of the crisis and the crisis type. For example, ATF represents Automatically labeled data for Texas Flood event while MGC represents Manually labeled data for Glasgow Crash event. Paid workers labeled the blue collections while the orange collections are unlabeled tweets retrieved by tweets IDs available in CrisisNLP. In contrast, the collections in green are the labeled data generated by our framework (CrisisLA).

Collection	# related tweets	# not related tweets	Total # tweets
Bohol Earthquake (MBE)	969	30	999
Queensland Floods (MQF)	919	280	1199
Colorado Floods (MCoF)	924	74	998
Manila Floods (MMF)	920	79	999
Alberta Floods (MAF)	982	17	999
Yolanda Typhoon (MYT)	939	108	1047
Sandy Typhoon (MST)	1581	429	2010
Oklahoma Typhoon (MOkT)	1769	241	2010
Nepal Earthquake (MNE)	2839	177	3016
Chile Earthquake (MChE)	1648	364	2013
California Earthquake (MCE)	169	13	182
Pakistan Earthquake (MPE)	1676	336	2012
India Floods (MIF)	1500	502	2002
Pakistan Floods (MPF)	1985	27	2012
Hagupit Typhoon (MHT)	1779	233	2012
Pam Typhoon (MPT)	1515	497	2012
Odile Typhoon (MOT)	178	4	182
Pakistan Earthquake (UPE)	-	-	35954
Pakistan Floods (UPF)	-	-	69521
Hagupit Typhoon (UHT)	-	-	26588
Indonesia Earthquake (AIE)	1900	300	2200
Texas Floods (ATF)	1100	400	1500
Sunda Strait Typhoon (AIT)	2000	400	1600

4 EXPERIMENTS

4.1 Datasets

We use specific collections from CrisisNLP [39], CrisisLex26 [19] and CrisisLexT26 [40] datasets to evaluate our framework (shown in Table 2). The datasets are labeled by paid workers based on either their relatedness to a given crisis event (CrisisLexT26 and CrisisLex26) or their corresponding informative class (CrisisNLP) (e.g., affected individuals, donations and volunteering, infrastructure and utilities, sympathy and support, other useful information and not related). However, for CrisisNLP data only, we relabel the available tweets into two classes: related and not related to a given crisis event. First, we combine all the tweets containing similar information such as "Personal updates" and "Affected individuals." Then, we relabel all the tweets from all the four classes except "not related" to a related class. "Not applicable" and other unclear labels have been discarded. We also eliminate the non-English tweets as our main goal is to build a reliable model for English tweets only, although it may be then transferred to other languages. To fulfil step four in our framework, we collected unlabeled tweets using Twitter API from three crisis events from three different crisis types: Texas Floods, Indonesia Earthquake and Sunda Strait tsunami (Typhoon). 4351 unlabeled data were collected for Texas Floods. Texas Floods data was crawled for five days from October 16 to 20, 2018, by using the hashtag "#flood" and geolocation information of Texas. We used the same methodology to collect 3989 unlabeled data of the Indonesia Earthquake event. However, we used the hashtag "#earthquake" and the geolocation information of Indonesia. The data was crawled for only one day, October 16, 2018, started from 16:00 to 23:59. Sunda Strait tsunami (Typhoon) was one of the strongest natural disasters that occurred in 2018 in Indonesia. The typhoon killed at least 426 people and 14060 were injured. We collected 145921 unlabeled tweets using the hashtag "#tsunami" and the geolocation of Sunda Strait. The data were crawled for an entire day on December 23, 2018.

Crisis-related Automatically Labeled dataset (CrisisAL) has been built using our framework. The created dataset includes automatically labeled tweets from three new crisis events. To train the model, we randomly select examples from these collections to approximately match the number of tweets of the other similar datasets from the same crisis type.

4.2 Evaluation Procedure

To answer our research questions, we ran two groups of experiments: The first group aimed at answering the first question, where we investigate the quality of the labeled data generated by our framework comparing to the manually labeled data from the same event. To do this, we conduct two sub experiments for each crisis

type: with manually labeled data and with automatically labeled data from a given disaster event. For example, in Earthquake crisis type, we train the model with all the manually labeled data, including Pakistan Earthquake (MPE). In the second experiment, MPE was replaced with the automatically labeled data related to Pakistan Earthquake (APE) to train the classifier. The second group of experiments aimed at showing the effectiveness of incorporating recent labeled data generated by our framework on the performance of the tweet classification model for crisis response. In these experiments, we compare three labeling methods which give labels to the unlabeled tweets from the recent events: Pseudo Labeling (PL) where similar manually labeled collections from the same crisis type were used to pretrained a model to be then used to label the recent unlabeled data (similar to [14]); Distant Supervision-based framework (DS) where our framework was used; and DS without FrameNet (DS-F) where step three was removed. Regarding the training data, we directly mixed the automatically labeled data with the available human-labeled data to train the tweet classifier. We also reported the classifier performance when trained using the original manually labeled data without incorporating the new labeled data (OG) to be used as our baseline.

It is worth noting that we used the same experimental setup, except the training data and the labeling method, for all the classifiers. To eliminate the noise in the tweets, we removed all the emojis, HTTP addresses, numbers, hashtags, user mentions and punctuations. Then all the examples were converted to lowercase and split into tokens to be passed to the model. We used Bidirectional Long Short-Term Memory (Bi-LSTM) [41] with a Maxpooling and 100-dimensional Global Vector Embeddings (GloVe) [42] as a pretrained word embedding since it is currently the best reported deep learning architecture for tweet classification for crisis response [43]. We used the same pre-processing and settings for all the experiments, and we repeated every experiment 30 times and took the mean as the final score.

5 RESULTS AND DISCUSSION

5.1 Quality of the Produced Data

As shown in Table 3, using APE instead of MPE with training datasets from other Earthquake events to classify the tweets in MChE, MBE, MCE and MNE datasets slightly drops the performance in F1 score by 1.2%, 2.8%, 0.5% and 1.2% respectively. Similar results are presented in Table 4 for Floods crisis type data, where the maximum drop is 4.2% on the MAF dataset. F1 scores displayed in Table 5 on four Typhoon event datasets determine minor decline in the model performance when using AHT instead of MHT in training them along with hand- labeled Typhoon events data. However, this is not the case for MST. One possible reason is that MST is very similar to one of the Typhoon (Hurricane) events in the previous training data.

In general, and to answer our first research question, it can be said that substituting automatically labeled data produced by our framework with manually labeled data from the same crisis event in training tweet classifiers for disaster response has a minimal impact on the classifier's performance for the three crisis types (< 5%). This is due to the noise (mislabeling problem) in the produced data. These results demonstrate that data annotated by our framework can be used when no hand-labeled data is available for new disaster events because they have similar quality levels. This finding can be considered as a good outcome; hence manually labeling new data from multiple events requires a lot of time, money and effort compared to the automatically generated data.

5.2 Effect of adding recent data

As can be seen in Table 6, DS reports the best labeling method for two Earthquake crisis datasets (MChE and MPE) with a maximum improvement of 2.1% in F1 score. On the other hand, the performance does not improve for MBE, MCE and MNE datasets. In Table 7, for Flood crisis datasets, DS is the best labeling method when tested on MAF, MIF, MCF and MQF datasets while PL is better than the other methods in the remain two datasets. However, the improvements in F1 score are very minor. For Typhoon crisis datasets, in Table 8, the classifier performance improves when using DS as the labeling method on three out of five datasets (MYT, MPT and MHT). After analyzing the data, we observe that more than seven keywords from the top K keyword list appear in MPE and MIF datasets which helps in providing new keywords from FrameNet to the training data while only one keyword occurs in MChE dataset with more than 30 new keywords driven from the external knowledge-base. These new keywords assist in recognizing related tweets that would not be identified by the old keyword list. And since we only label tweets with two keywords, different (new) relations may emerge using these new keywords. On the other hand, in the case of the limited number of new keywords driven from FrameNet, adding data from new crisis event does not improve the model performance regardless of the number of the top K keywords appear in the test data especially if the training and the testing data are dissimilar. If the train and the test data are similar and the number of matched keywords is low, then, PL is the predicted best labeling method. To answer the second research question, we can say that there is no significant improvement in the model performance when adding the automatically labeled data produced by neither any of the three labeling methods (DS, PL or DS-FN) to the original manually labeled training data. Generally, results indicate that DS is the best labeling method if new driven keywords from FrameNet exist in the test data, especially if the similarity between the test and train data is low. However, more future experiments are needed to examine the effectiveness of adding more than one crisis event's data at the same time and to test the models on recent 2019 data.

6 CONCLUSION AND FUTURE WORK

In this paper, we investigate the application of distant supervision in generating automatically labeled tweets from new crisis events to overcome the problem of reduced generalization levels of the current crisis-related classifiers when tested on tweets from unseen events. Reducing the generalization errors leads to a more reliable system to be used by humanitarian organizations to help people in need during crises. The results show the effectiveness of our distant supervision- based framework in producing labeled training tweets from new crisis events to train the classifier, especially when no manually labeled data is available for the given crisis event. Substituting generated annotated data instead of manually labeled

Table 3: Results in F1 score for the first experiment group in Earthquake crisis data. E is all the available manually labeled Earthquake crisis datasets excluding MPE.

Train/ Test	MChE	MBE	MCE	MNE
E+MPE	0.8044	0.9335	0.8941	0.9168
E+APE	0.7882	0.9052	0.8890	0.9043

Table 4: Results in F1 score for the first experiment group in Flood crisis data. F is all the available manually labeled Floods crisis datasets excluding MPF.

Train/ Test	MIF	MCF	MAF	MMF	MQF
F+MPF	0.7389	0.9224	0.9619	0.8987	0.7855
F+APF	0.7100	0.9170	0.9190	0.8881	0.7496

Table 5: Results in F1 score for the first experiment group in Typhoon crisis data. T is all the available manually labeled Typhoon crisis datasets excluding MHT.

Train/ Test	MOkT	MYT	MST	MOT	MPT
T+MHT	0.8418	0.8781	0.8023	0.9618	0.7072
T+AHT	0.8067	0.8372	0.7242	0.9126	0.6955

Table 6: Results in F1 score for the second experiment group in Earthquake.

Model/ Test	MPE	MNE	MCE	MBE	MChE
OG	0.7915	**0.9068**	**0.8921**	**0.8876**	0.8356
PL	0.7903	0.9045	0.8842	**0.8877**	0.8302
DS	**0.7940**	0.8875	0.8769	0.8863	**0.8566**
DS-F	0.7913	0.9026	0.8780	0.8855	0.8581

data in training tweet classifiers for disaster response has a small impact on the performance. The performance dropped for less than 5% on 13 out of 14 datasets from different locations and crisis types. This indicates that the generated data has a competitive quality compared to the manually labeled data with less effort, time and money. Results also suggest that our proposed framework is the best labeling method when the test and the train data are dissimilar. This is because it can recognize the related tweets in

Table 7: Results in F1 score for the second experiment group in Flood.

Model/ Test	MPF	MQF	MCF	MIF	MAF
OG	0.962	0.839	0.917	0.764	0.916
PL	**0.968**	0.836	0.921	0.762	0.917
DS	0.960	**0.840**	**0.922**	**0.767**	**0.925**
DS-F	0.966	**0.840**	0.920	0.761	0.920

Table 8: Results in F1 score for the second experiment group in Typhoon.

Model/ Test	MHT	MPT	MYT	MOT	MOkT	MST
OG	0.881	0.827	0.901	0.961	**0.793**	**0.708**
PL	0.881	0.825	0.897	**0.962**	0.787	0.703
DS	**0.882**	**0.829**	**0.9117**	0.957	0.759	0.699
DS-F	**0.883**	0.828	0.910	0.960	0.777	0.702

the test data that include new keywords retrieved from FrameNet and do not exist in the original training data. However, adding these tweets to the previously available human-labeled tweets in training the classifiers did not have significant improvements in the performance. More future experiments are needed to examine the effectiveness of adding automatically annotated data from two crisis events instead of one to cover the gap in the tweets number between the manually and the automatically labeled data. In addition, we intend to use our framework to generate datasets for crisis types that have only one or two manually labeled collections, such as building collapse.

REFERENCES

[1] Vieweg, S. E. (2012). Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications (Doctoral dissertation, University of Colorado at Boulder).

[2] Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. In Eleventh International AAAI Conference on Web and Social Media.

[3] Qu, Y., Huang, C., Zhang, P., & Zhang, J. 2011. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. In Proceedings of the ACM 2011 conference on Computer supported cooperative work (pp. 25- 34). ACM.

[4] Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. 2010. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In Proceedings of the 2010 ACM conference on Computer supported cooperative work (pp. 241-250). ACM.

[5] Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1079-1088). ACM.

[6] Thomson, R., Ito, N., Suda, H., Lin, F., Liu, Y., Hayasaka, R., ... & Wang, Z. 2012. Trusting tweets: The Fukushima disaster and information source credibility on Twitter. In Proceedings of the 9th International ISCRAM Conference (pp. 1-10). Vancouver: Simon Fraser University.

[7] Cho, S. E., Jung, K., & Park, H. W. 2013. Social media use during Japan's 2011 earthquake: how Twitter transforms the locus of crisis communication. Media International Australia, 149(1), 28-40.

[8] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12), 2009.

[9] Tapia, A. H., & Moore, K. (2014). Good enough is good enough: Overcoming disaster response organizations' slow social media data adoption. Computer Supported Cooperative Work (CSCW), 23(4-6), 483-512.

[10] Caragea, C., Silvescu, A., & Tapia, A. H. (2016, May). Identifying informative messages in disaster events using convolutional neural networks. In International Conference on Information Systems for Crisis Response and Management (pp. 137-147).

[11] Li, H., Caragea, D., Caragea, C., & Herndon, N. (2018). Disaster response aided by tweet classification with a domain adaptation approach. Journal of Contingencies and Crisis Management, 26(1), 16-27.

[12] Palen, L., & Anderson, K. M. (2016). Crisis informatics—New data for extraordinary times. Science, 353(6296), 224-225.

[13] ALRashdi, R., & O'Keefe, S. (2019, November). Robust Domain Adaptation Approach for Tweet Classification for Crisis Response. In International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning (pp. 124-134). Springer, Cham.

[14] Win, S. S. M., & Aung, T. N. (2018). Automated Text Annotation for Social Media Data during Natural Disasters. Advances in Science, Technology and Engineering Systems Journal, 3(2), 119-27.

[15] Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009, August). Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2(pp. 1003-1011). Association for Computational Linguistics.

[16] Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The berkeley framenet project. In Proceedings of the 17th international conference on Computational linguistics-Volume 1 (pp. 86-90). Association for Computational Linguistics.

[17] To, H., Agrawal, S., Kim, S. H., & Shahabi, C. (2017, April). On identifying disaster-related tweets: Matching-based or learning-based?. In 2017 IEEE Third International Conference on Multimedia Big Data (BigMM) (pp. 330-337). IEEE.

[18] Kumar, S., Barbier, G., Abbasi, M. A., & Liu, H. (2011, July). Tweettracker: An analysis tool for humanitarian and disaster relief. In Fifth international AAAI conference on weblogs and social media.

[19] Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014, May). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In Eighth International AAAI Conference on Weblogs and Social Media.

[20] Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. Sociological research online, 18(3), 1-11.

[21] Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E., & Laredo, J. A. (2013). CrisisTracker: Crowdsourced social media curation for disaster awareness. IBM Journal of Research and Development, 57(5), 4-1.

[22] Doshi, T., Marriott, E., & Patel, J. (2017). CS224N Final Project: Detecting Key Needs in Crisis.

[23] Win, S. S. M., & Aung, T. N. (2017, May). Target oriented tweets monitoring system during natural disasters. In2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) (pp. 143-148). IEEE.

[24] Neppalli, V. K., Caragea, C., & Caragea, D. (2018). Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters. In ISCRAM.

[25] Zeng, Y., Feng, Y., Ma, R., Wang, Z., Yan, R., Shi, C., & Zhao, D. (2018, April). Scale Up Event Extraction Learning via Automatic Training Data Generation. In Thirty-Second AAAI Conference on Artificial Intelligence.

[26] Chen, Y., Liu, S., Zhang, X., Liu, K., & Zhao, J. (2017, July). Automatically labeled data generation for large scale event extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 409-419).

[27] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12), 2009.

[28] Mohammed, S., Ghelani, N., & Lin, J. (2017). Distant Supervision for Topic Classification of Tweets in Curated Streams. arXiv preprint arXiv:1704.06726.

[29] Magdy, W., Sajjad, H., El-Ganainy, T., & Sebastiani, F. (2015, April). Distant supervision for tweet classification using youtube labels. In Ninth International AAAI Conference on Web and Social Media.

[30] Qu, J., Ouyang, D., Hua, W., Ye, Y., & Li, X. (2018). Distant supervision for neural relation extraction integrated with word attention and property features. Neural Networks, 100, 59-69.

[31] Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29.

[32] Riedel, S., Yao, L., & McCallum, A. (2010, September). Modeling relations and their mentions without labeled text. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 148-163). Springer, Berlin, Heidelberg.

[33] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., & Weld, D. S. (2011, June). Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Volume 1 (pp. 541-550). Association for Computational Linguistics.

[34] Zheng, W., & Blake, C. (2015). Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles. Journal of biomedical informatics, 57, 134-144.

[35] Li, G., Wu, C., & Vijay-Shanker, K. (2017, August). Noise reduction methods for distantly supervised biomedical relation extraction. In BioNLP 2017 (pp. 184-193).

[36] Su, P., Li, G., Wu, C., & Vijay-Shanker, K. (2019). Using distant supervision to augment manually annotated data for relation extraction. BioRxiv, 626226.

[37] Takamatsu, S., Sato, I., & Nakagawa, H. (2012, July). Reducing wrong labels in distant supervision for relation extraction. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 721-729). Association for Computational Linguistics.

[38] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.

[39] Imran, M., Mitra, P., & Castillo, C. (2016, May). Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (pp. 1638-1643).

[40] Olteanu, A., Vieweg, S., & Castillo, C. (2015, February). What to expect when the unexpected happens: Social media communications across crises. In Proceedings of the 18th ACM conference on computer supported cooperative work & social computing (pp. 994-1009). ACM.

[41] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[42] Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[43] </number>ALRashdi, R., & O'Keefe, S. (2018, October). Deep Learning and Word Embeddings for Tweet Classification for Crisis Response. In The 3rd National Computing Colleges Conference. Abha.